Deep web

The **deep web**,[1] **invisible web**,[2] or **hidden web**[3] are parts of the World Wide Web whose contents are not indexed by standard web search engines. The opposite term to the deep web is the surface web, which is accessible to anyone using the Internet.[4] Computer scientist Michael K. Bergman is credited with coining the term *deep web* in 2001 as a search indexing term.[5]

The content of the deep web is hidden behind HTTP forms[6][7] and includes many very common uses such as web mail, online banking, and services that users must pay for, and which are protected by paywalls, such as video on demand and some online magazines and newspapers.

The content of the deep web can be located and accessed by a direct URL or IP address, and may require a password or other security access past the public website page.

Contents

Terminology
Non-indexed content
Indexing methods
Content types
See also
References
Further reading
External links

Terminology

The first conflation of the terms "deep web" with "dark web" came about in 2009 when deep web search terminology was discussed together with illegal activities taking place on the Freenet and darknet.[8]

Since then, after their use in the media's reporting on the Silk Road, many[9][10] people and media outlets have taken to using 'deep web' synonymously with the dark web or darknet, a comparison some reject as inaccurate[11] and consequently has become an ongoing source of confusion.[12] *Wired* reporters Kim Zetter[13] and Andy Greenberg[14] recommend the terms be used in distinct fashions. While the deep web is a reference to any site that cannot be accessed through a traditional search engine, the dark web is a portion of the deep web that has been intentionally hidden and is inaccessible through standard browsers and methods.[15][16][17][18][19]

Non-indexed content

Bergman, in a paper on the deep web published in *The Journal of Electronic Publishing*, mentioned that Jill Ellsworth used the term *Invisible Web* in 1994 to refer to websites that were not registered with any search engine.[20] Bergman cited a January 1996 article by Frank Garcia:[21]

> It would be a site that's possibly reasonably designed, but they didn't bother to register it with any of the search engines. So, no one can find them! You're hidden. I call that the invisible Web.

Another early use of the term *Invisible Web* was by Bruce Mount and Matthew B. Koll of Personal Library Software, in a description of the #1 Deep Web tool found in a December 1996 press release.[22]

The first use of the specific term *deep web*, now generally accepted, occurred in the aforementioned 2001 Bergman study.[20]

Indexing methods

Methods that prevent web pages from being indexed by traditional search engines may be categorized as one or more of the following:

1. **Contextual web**: pages with content varying for different access contexts (e.g., ranges of client IP addresses or previous navigation sequence).
2. **Dynamic content**: dynamic pages, which are returned in response to a submitted query or accessed only through a form, especially if open-domain input elements (such as text fields) are used; such fields are hard to navigate without domain knowledge.
3. **Limited access content**: sites that limit access to their pages in a technical way (e.g., using the Robots Exclusion Standard or CAPTCHAs, or no-store directive, which prohibit search engines from browsing them and creating cached copies).[23]
4. **Non-HTML/text content**: textual content encoded in multimedia (image or video) files or specific file formats not handled by search engines.
5. **Private web**: sites that require registration and login (password-protected resources).
6. **Scripted content**: pages that are only accessible through links produced by JavaScript as well as content dynamically downloaded from Web servers via Flash or Ajax solutions.
7. **Software**: certain content is intentionally hidden from the regular Internet, accessible only with special software, such as Tor, I2P, or other darknet software. For example, Tor allows users to access websites using the .onion server address anonymously, hiding their IP address.
8. **Unlinked content**: pages which are not linked to by other pages, which may prevent web crawling programs from accessing the content. This content is referred to as pages without backlinks (also known as inlinks). Also, search engines do not always detect all backlinks from searched web pages.
9. **Web archives**: Web archival services such as the Wayback Machine enable users to see archived versions of web pages across time, including websites which have become inaccessible, and are not indexed by search engines such as Google.[24]

Content types

While it is not always possible to directly discover a specific web server's content so that it may be indexed, a site potentially can be accessed indirectly (due to computer vulnerabilities).

To discover content on the web, search engines use web crawlers that follow hyperlinks through known protocol virtual port numbers. This technique is ideal for discovering content on the surface web but is often ineffective at finding deep web content. For example, these crawlers do not attempt to find dynamic pages that are the result of database queries due to the indeterminate number of queries that are possible.[5] It has been noted that this can be (partially) overcome by providing links to query results, but this could unintentionally inflate the popularity for a member of the deep web.

DeepPeep, Intute, Deep Web Technologies, Scirus, and Ahmia.fi are a few search engines that have accessed the deep web. Intute ran out of funding and is now a temporary static archive as of July 2011.[25] Scirus retired near the end of January 2013.[26]

Researchers have been exploring how the deep web can be crawled in an automatic fashion, including content that can be accessed only by special software such as Tor. In 2001, Sriram Raghavan and Hector Garcia-Molina (Stanford Computer Science Department, Stanford University)[27][28] presented an architectural model for a hidden-Web crawler that used key terms provided by users or collected from the query interfaces to query a Web form and crawl the Deep Web content. Alexandros Ntoulas, Petros Zerfos, and Junghoo Cho of UCLA created a hidden-Web crawler that automatically generated meaningful queries to issue against search forms.[29] Several form query languages (e.g., DEQUEL[30]) have been proposed that, besides issuing a query, also allow extraction of structured data from result pages. Another effort is DeepPeep, a project of the University of Utah sponsored by the National Science Foundation, which gathered hidden-web sources (web forms) in different domains based on novel focused crawler techniques.[31][32]

Commercial search engines have begun exploring alternative methods to crawl the deep web. The Sitemap Protocol (first developed, and introduced by Google in 2005) and OAI-PMH are mechanisms that allow search engines and other interested parties to discover deep web resources on particular web servers. Both mechanisms allow web servers to advertise the URLs that are accessible on them, thereby allowing automatic discovery of resources that are not directly linked to the surface web. Google's deep web surfacing syster computes submissions for each HTML form and adds the resulting HTML pages into the Google search engine index. The surfaced results account for a thousand queries per second to deep web content.[33] In this system, the pre-computation of submissions is done using three algorithms:

1. selecting input values for text search inputs that accept keywords,
2. identifying inputs which accept only values of a specific type (e.g., date) and
3. selecting a small number of input combinations that generate URLs suitable for inclusion into the web search index.

In 2008, to facilitate users of Tor hidden services in their access and search of a hidden .onion suffix, Aaron Swartz designed Tor2web—a proxy application able to provide access by means of common web browsers.[34] Using this application, deep web links appear as a random string of letters followed by the .onion TLD.

See also

- Deep linking
- Gopher protocol
- DARPA's Memex program

References

1. Hamilton, Nigel (2003). "The Mechanics of a Deep Net Metasearch Engine" (http://www.iadisportal.org/digital-library/the-mechanics-of-a-deep-net-metasearch-engine) In Isaías, Pedro; Palma dos Reis, António (eds.) *Proceedings of the IADIS International Conference on e-Society* pp. 1034–6. CiteSeerX 10.1.1.90.5847 (https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.90.5847) ISBN 972-98947-0-1.
2. Devine, Jane; Egger-Sider, Francine (July 2004). "Beyond google: the invisible web in the academic library" *The Journal of Academic Librarianship* **30** (4): 265–269. doi:10.1016/j.acalib.2004.04.010 (https://doi.org/10.1016%2Fj.acalib.2004.04.010).
3. Raghavan, Sriram; Garcia-Molina, Hector (September 11–14, 2001). "Crawling the Hidden Web" (http://ilpubs.stanford.edu:8090/725/). *27th International Conference on Very Large Data Bases*.
4. "Surface Web" (https://www.computerhope.com/jargon/s/surface-web.htm). Computer Hope. Retrieved June 20, 2018.
5. Wright, Alex (February 22, 2009). "Exploring a 'Deep Web' That Google Can't Grasp" (https://www.nytimes.com/2009/02/23/technology/internet/23search.html?th&emc=th) *The New York Times*. Retrieved February 23, 2009.
6. Madhavan, J., Ko, D., Kot, Ł., Ganapathy, V., Rasmussen, A., & Halevy, A. (2008). Google's deep web crawl. Proceedings of the VLDB Endowment, 1(2), 1241–52.
7. Shedden, Sam (June 8, 2014). "How Do You Want Me to Do It? Does It Have to Look like an Accident? – an Assassin Selling a Hit on the Net; Revealed Inside the Deep Web" (https://www.questia.com/article/1G1-370513892/how-do-you-want-me-to-do-it-does-it-have-to-look) *Sunday Mail*. Retrieved May 5, 2017 – via Questia.
8. Beckett, Andy (November 26, 2009). "The dark side of the internet" (https://www.theguardian.com/technology/2009/nov/26/dark-side-internet-freenet) Retrieved August 9, 2015.
9. Daily Mail Reporter (October 11, 2013). "The disturbing world of the Deep Web, where contract killers and drug dealers ply their trade on the internet" (http://www.dailymail.co.uk/news/article-2454735/The-disturbing-world-Deep-Web-contract-killers-drug-dealers-ply-trade-internet.html). Retrieved May 25, 2015.
10. "NASA is indexing the 'Deep Web' to show mankind what Google won't" (http://fusion.net/story/145885/nasa-is-indexing-the-deep-web-to-show-mankind-what-google-wont/) Fusion.
11. "Clearing Up Confusion – Deep Web vs. Dark Web" (http://www.brightplanet.com/2014/03/clearing-confusion-deep-web-vs-dark-web/). BrightPlanet. March 27, 2014.
12. Solomon, Jane (May 6, 2015). "The Deep Web vs. The Dark Web" (http://blog.dictionary.com/dark-web/). Retrieved May 26, 2015.

13. NPR Staff (May 25, 2014). "Going Dark: The Internet Behind The Internet" (https://www.npr.org/sections/alltechconsidered/2014/05/25/315821415/going-dark-the-internet-behind-the-internet). Retrieved May 29, 2015.
14. Greenberg, Andy (November 19, 2014). "Hacker Lexicon: What Is the Dark Web?" (https://www.wired.com/2014/11/hacker-lexicon-whats-dark-web/). Retrieved June 6, 2015.
15. "The Impact of the Dark Web on Internet Governance and Cyber Security" (https://www.ourinternet.org/sites/default/files/publications/GCIG_Paper_No6.pdf) (PDF). Retrieved January 15, 2017.
16. Lam, Kwok-Yan; Chi, Chi-Hung; Qing, Sihan (November 23, 2016). *Information and Communications Security: 18th International Conference, ICICS 2016, Singapore, Singapore, November 29 – December 2, 2016, Proceedings* (https://books.google.com/books?id=uraVDQAAQBAJ&pg=PR174). Springer. ISBN 9783319500119. Retrieved January 15, 2017.
17. "The Deep Web vs. The Dark Web | Dictionary.com Blog" (http://blog.dictionary.com/dark-web/). Dictionary Blog. May 6, 2015. Retrieved January 15, 2017.
18. Akhgar, Babak; Bayerl, P. Saskia; Sampson, Fraser (January 1, 2017). *Open Source Intelligence Investigation: From Strategy to Implementation* (https://books.google.com/books?id=39zTDQAAQBAJ&pg=PR112). Springer. ISBN 9783319476711. Retrieved January 15, 2017.
19. "What is the dark web and who uses it?" (https://www.theglobeandmail.com/technology/tech-news/what-is-the-dark-web-and-who-uses-it/article26026082/) *The Globe and Mail*. Retrieved January 15, 2017.
20. Bergman, Michael K (August 2001). "The Deep Web: Surfacing Hidden Value" (http://quod.lib.umich.edu/cgi/t/text/text-idx?c=jep;view=text;rgn=main;idno=3336451.0007.104) *The Journal of Electronic Publishing* **7** (1). doi:10.3998/3336451.0007.104 (https://doi.org/10.3998%2F3336451.0007.104)
21. Garcia, Frank (January 1996). "Business and Marketing on the Internet" (https://web.archive.org/web/19961205083117/http://tcp.ca/Jan96/BusandMark.html) *Masthead*. **15** (1). Archived from the original (http://tcp.ca/Jan96/BusandMark.html) on December 5, 1996. Retrieved February 24, 2009.
22. @1 started with 5.7 terabytes of content, estimated to be 30 times the size of the nascent World Wide Web; PLS was acquired by AOL in 1998 and @1 was abandoned. "PLS introduces AT1, the first 'second generation' Internet search service" (https://web.archive.org/web/19971021232057/http://www.pls.com/news/pr961212_at1.html) (Press release). Personal Library Software. December 1996. Archived from the original (https://www.pls.com/news/pr961212_at1.html) on October 21, 1997. Retrieved February 24, 2009.
23. "Hypertext Transfer Protocol (HTTP/1.1): Caching" (http://tools.ietf.org/html/rfc7234#section-5.2.2.3). Internet Engineering Task Force. 2014. Retrieved July 30, 2014.
24. Wiener-Bronner, Danielle (June 10, 2015). "NASA is indexing the 'Deep Web' to show mankind what Google won't" (http://fusion.net/story/145885/nasa-is-indexing-the-deep-web-to-show-mankind-what-google-wont/). Fusion. Retrieved June 27, 2015. "There are other simpler versions of Memex already available. "If you've ever used the Internet Archive's Wayback Machine", whichgives you past versions of a website not accessible through Google, then you've technically searched the Deep Web, said Chris Mattmann."
25. "Intute FAQ" (http://www.intute.ac.uk/faq.html). Retrieved October 13, 2012.
26. "Elsevier to Retire Popular Science Search Engine" (http://library.bldrdoc.gov/newsarc/201312.html). *library.bldrdoc.gov*. December 2013. Retrieved June 22, 2015. "by end of January 2014, Elsevier will be discontinuing Scirus, its free science search engine. Scirus has been a wide-ranging research tool, with over 575 million items indexed for searching, including webpages, pre-print articles, patents, and repositories."
27. Sriram Raghavan (http://www.almaden.ibm.com/cs/people/rsriram/); Garcia-Molina, Hector (2000). "Crawling the Hidden Web" (http://ilpubs.stanford.edu:8090/456/1/2000-36.pdf) (PDF). Stanford Digital Libraries Technical Report. Retrieved December 27, 2008.
28. Raghavan, Sriram; Garcia-Molina, Hector (2001). "Crawling the Hidden Web" (http://www.dia.uniroma3.it/~vldbproc/017_129.pdf) (PDF). *Proceedings of the 27th International Conference on Very Large Data Bases (VLDB)* pp. 129–38.
29. Alexandros, Ntoulas; Zerfos, Petros; Cho, Junghoo (2005). "Downloading Hidden Web Content" (http://oak.cs.ucla.edu/~cho/papers/ntoulas-hidden.pdf) (PDF). UCLA Computer Science. Retrieved February 24, 2009.
30. Shestakov, Denis; Bhowmick, Sourav S.; Lim, Ee-Peng (2005). "DEQUE: Querying the Deep Web" (http://www.inf.ufsc.br/~r.mello/deepWeb/querying/DKE2005-Sourav.pdf) (PDF). *Data & Knowledge Engineering* **52** (3): 273–311.
31. Barbosa, Luciano; Freire, Juliana (2007). "An Adaptive Crawler for Locating Hidden-Web Entry Points" (http://www.cs.utah.edu/~lbarbosa/publications/ache-www2007.pdf) (PDF). WWW Conference 2007. Retrieved March 20, 2009.

32. Barbosa, Luciano; Freire, Juliana (2005). "Searching for Hidden-Web Databases" (http://www.cs.utah.edu/~lbarbosa/publications/webdb2005.pdf) (PDF). WebDB 2005. Retrieved March 20, 2009.
33. Madhavan, Jayant; Ko, David; Kot, Łucja; Ganapathy, Vignesh; Rasmussen, Alex; Halevy, Alon (2008). "Google's Deep-Web Crawl" (https://homes.cs.washington.edu/~alon/files/vldb08deepweb.pdf) (PDF). VLDB Endowment, ACM. Retrieved April 17, 2009.
34. Aaron, Swartz. "In Defense of Anonymity" (http://www.aaronsw.com/weblog/tor2web) Retrieved February 4, 2014.

Further reading

- Barker, Joe (January 2004). "Invisible Web: What it is, Why it exists, How to find it, and its inherent ambiguity". University of California, Berkeley Teaching Library Internet Workshops. Archived from the original on July 29, 2005. Retrieved July 26, 2011..
- Basu, Saikat (March 14, 2010). "10 Search Engines to Explore the Invisible Web". MakeUseOf.com.
- Ozkan, Akin (November 2014). "Deep Web /Derin İnternet"..
- Gruchawka, Steve (June 2006). "How-To Guide to the Deep Web"..
- Hamilton, Nigel (2003). "The Mechanics of a Deep Net Metasearch Engine" 12th World Wide Web Conference.
- He, Bin; Chang, Kevin Chen-Chuan (2003). "Statistical Schema Matching across Web Query Interfaces" (PDF). *Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data*. Archived from the original (PDF) on July 20, 2011.
- Howell O'Neill, Patrick (October 2013). "How to search the Deep Web". The Daily Dot.
- Ipeirotis, Panagiotis G.; Gravano, Luis; Sahami, Mehran (2001). "Probe, Count, and Classify: Categorizing Hidden-Web Databases" (PDF). *Proceedings of the 2001 ACM SIGMOD International Conference on Management of Data*. pp. 67–78. Archived from the original (PDF) on September 12, 2006 Retrieved September 26, 2006.
- King, John D.; Li, Yuefeng; Tao, Daniel; Nayak, Richi (November 2007). "Mining World Knowledge for Analysis of Search Engine Content" (PDF). *Web Intelligence and Agent Systems: An International Journal*. **5** (3): 233–53. Archived from the original (PDF) on December 3, 2008 Retrieved July 26, 2011.
- McCown, Frank; Liu, Xiaoming; Nelson, Michael L.; Zubair, Mohammad (March – April 2006). "Search Engine Coverage of the OAI-PMH Corpus" (PDF). *IEEE Internet Computing* **10** (2): 66–73. doi:10.1109/MIC.2006.41
- Price, Gary; Sherman, Chris (July 2001). *The Invisible Web: Uncovering Information Sources Search Engines Can't See*. CyberAge Books. ISBN 978-0-910965-51-4
- Shestakov, Denis (June 2008). *Search Interfaces on the Web: Querying and Characterizing*. TUCS Doctoral Dissertations 104, University of Turku
- Whoriskey, Peter (December 11, 2008). "Firms Push for a More Searchable Federal Web". *The Washington Post*. p. D01..
- Wright, Alex (March 2004). "In Search of the Deep Web". Salon. Archived from the original on March 9, 2007.
- Scientists, Naked (December 2014). "The Internet: the good, the bad and the ugly – In-depth exploration of the Internet and the Dark Web by Cambridge University's Naked Scientists" (Podcast).

External links

Retrieved from 'https://en.wikipedia.org/w/index.php?title=Deep_web&oldid=896914536'

This page was last edited on 13 May 2019, at 16:49 (UTC).

Text is available under the Creative Commons Attribution-ShareAlike License; additional terms may apply. By using this site, you agree to the Terms of Use and Privacy Policy. Wikipedia® is a registered trademark of the Wikimedia Foundation, Inc., a non-profit organization.